Landscape of exitrons in gastric cancer

Yihao Zhang,[a,b] Gengtai Ye,[a,b] Qingbin Yang,[a,b] Boyang Zheng,[a,b] Guofan Zhang,[a,b] Yanfeng Hu,[a,b] Jiang Yu,[a,b] and Guoxin Li[a,b]*

[a]Department of General Surgery, Nanfang Hospital, Southern Medical University, Guangdong Provincial Engineering Technology Research Center of Minimally Invasive Surgery, Guangzhou, Guangdong 510515, China
[b]Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Tumor, Guangzhou, Guangdong 510515, China

eBioMedicine 2022;84: 104272
Published online xxx
https://doi.org/10.1016/j.ebiom.2022.104272

Summary

Background Exitron is a new type of non-canonical alternative splicing. Accumulating evidence implies exitron may have pathological function and contribute to another source of anti-tumor immunogenicity in various cancers. Its role in gastric cancer remains poorly understood. Large-scale, multi-omics analysis could comprehensively characterize the landscape of exitrons in gastric cancer, reveal undiscovered mechanism and hopefully identify molecular biomarkers for predicting immunotherapy response.

Methods We collected datasets from five studies for analysis. RNA sequencing was used for exitron identification. Somatic mutations were detected by whole exome sequencing. Neopeptides were confirmed by proteome mass spectrometry.

Findings 42174 gastric cancer-specific exitrons (GCSEs) were identified in 632 patients. GCSEs were clinically relevant to gender, age, Lauren type, tumor stage and prognosis. Tissue specificity test and pathogenic exitron prediction revealed their unique functional impact. GCSEs were mutually exclusive with mutations and demonstrated both unique and complementary function against TP53 mutation in gastric cancer. We further established splicing regulatory network to reveal upstream regulation of exitron splicing. We also evaluated the immunogenicity and diagnostic potential of GCSEs. Evidence of GCSEs-derived neopeptide expression was validated by whole proteome mass spectrometry. PD-1 and Siglecs were significantly increased in high neoantigen load patients. But exitron-related biomarkers failed to predict immunotherapy response, possibly due to small sample size and insufficient sequencing depth.

Interpretation The present study provided a comprehensive multidimensional landscape of gastric cancer exitrons and underscores insights into underexplored mechanism in gastric cancer pathology.

Funding The Guangdong Provincial Key Laboratory of Precision Medicine for Gastroinstestinal Cancer (2020B121201004), the Guangdong Provincial Major Talents Project (No. 2019JC05Y361) and National Natural Science Foundation of China (grant number:82172960 and 81872013).

Keywords: RNA splicing; Exitron; Multi-omics; Gastric cancer

Introduction

Gastric cancer is the 3rd leading cause of cancer-related mortality, representing 8.2% of all deaths from cancer.[1] For the last decade, exploring genetic/epigenetic alteration has revealed mechanistic complexity in gastric cancer. Chronic infection of *H. pylori* and Epstein-Barr virus, mutated driver genes, driver gene amplification have been demonstrated to contribute to gastric cancer pathology.[2-4] Efforts have been made to develop targeted therapies but few have shown clinical benefit, which reflects the fact that pathological mechanism in gastric cancer remains largely unknown. Recently, accumulating evidence revealed an important role of alternative splicing in cancer. Disrupted alternative splicing is commonly observed in various cancers and can functionally drive oncogenic process in a driver gene mutation-complementary manner[5] or produce non-functioning transcriptional noise to promote anti-

*Corresponding author.
E-mail address: gzliguoxin@163.com (G. Li).

Research in Context

Evidence before this study

Non-canonical alternative splicing, exonic intron (exitron) in gastric cancer has been largely overlooked. Previous studies described exitron features in breast cancer, prostate cancer, and other cancer types. But the characteristics and function of exitrons in gastric cancer remained poorly understood. In this study, we comprehensively illustrated a multidimensional landscape of gastric cancer exitron, identified exitron-dependent machinery related to oncogenesis of gastric cancer, and evaluated the clinical application potential of gastric cancer exitrons in immunotherapy.

Added value of this study

Our study included transcriptome of 632 patients, whole exome of 481 patients and proteome of 77 patients to characterize gastric cancer exitrons. We presented in-depth analyses of exitrons splicing, somatic mutations, *in-silico* functional assay, pathogenic prediction, mutual exclusivity, upstream splicing regulatory network, neoantigen prediction, neopeptide validation, and immunotherapy response prediction.

Implications of all the available evidence

Pathological mechanism of gastric cancer remains largely unknown. Comprehensive understanding of gastric cancer exitrons could provide insight into discovering pathological mechanism, oncogenic candidates, diagnostic biomarkers and treatment targets. Our study characterized landscape of gastric cancer exitron, demonstrated its unique function and features in oncogenic process, and provided important information for genetic screening and functional validation of potential candidates in the future.

tumor immunogenicity by increasing neoantigen load.[6]

Exonic intron (in short, exitron) is a new type of non-canonical alternative splicing.[7] These cryptic introns could be spliced inside annotated exons, thus contain both splicing and coding potential. Exitron splicing has been observed in both tumor and normal samples.[7-9] In two pilot studies, exitrons recurrently spliced within tumor suppressor genes in breast cancer and prostate cancer.[7,8] In a pan-cancer analysis, 129406 exitrons were discovered in 9599 tumor samples.[9] Among 33 cancer types, ovarian cancer, esophageal cancer, gastric cancer are the top three tumors with highest exitron load. On the one hand, some exitrons could exhibit oncogenic potential. For example, exitron-spliced NEFH produced a progressive phenotype in prostate cancer cells, indicating its tumor suppressor role in prostate cancer. On the other hand, exitrons could increase immunogenicity as the expression and presentation of exitron-derived neoantigen was validated in ovarian cancer and breast cancer patients. Moreover, putative neoantigen burden could predict response to immune checkpoint inhibitor in clear cell renal cell carcinoma patients. These findings underscore the importance of investigating exitrons in cancer pathology.

Currently, a full picture about exitron splicing in gastric cancer is still lacking due to insufficient samples for comprehensive analysis. Meanwhile, it is worth noting that characteristics of exitron events in a particular cancer could be underrepresented in pan-cancer analysis.[10] In the present study, we illustrated a multidimensional comprehensive landscape of exitron splicing in gastric cancer. We characterized the biological and clinical feature of gastric cancer exitrons. The pathogenicity of gastric cancer exitrons was validated by tissue-specific test and pathogenicity prediction. We demonstrated unique functions of gastric cancer exitrons. Furthermore, we investigated mutual exclusivity of gastric cancer exitrons and somatic mutations and identified two groups of compensatory/synergistic exitrons against p53 mutations. To discover the upstream regulatory candidates of gastric cancer exitron splicing, we built molecular networks connecting expression/mutation profile of splicing factors with exitrons. Additionally, we confirmed the expression of gastric cancer-specific exitrons (GCSEs)-derived transcripts by mass spectrometry (MS) analysis. The expression of PD-1 and Siglecs family members were significantly increased in patients with high GCSEs-derived neoantigen load, indicating GCSEs were of value to clinical diagnosis and treatment prediction.

Methods

Clinical cohort and data usage

We systematically searched PubMed, GEO dataset, European Nucleotide Archive (ENA) to find studies including gastric cancer patient samples. No other limitation was applied for searching. Overall, five clinical cohorts from United States, Korea and China were included.[9,11-13] Five datasets contained transcriptome data. Three datasets contained whole exome data. One dataset contained whole proteome data. Information including age, gender, stage, Lauren pathology classification, overall survival time, follow-up status and immune checkpoint inhibitor response were collected from these datasets if available. In the present study, we mainly conducted 11 experiments using abovementioned data. Detailed dataset characteristics and data usage for each experiment can be found in Table S1.

Data quality control

Fastp 0.20.0[14] was used to perform quality assessment and reads filtering of raw FASTQ files in RNA-Seq and whole exome sequencing (WES) datasets. Briefly,

adapters were automatically detected by setting parameter *–detect_adapter_for_pe*. Reads with mean read quality less than 20 within a range of 4bp sliding window were filtered. The first 15bp in front of reads were trimmed.

Sequencing reads alignment

All RNA-Seq data were aligned to GENCODE hg38 reference genome using splice junction-aware software STAR 2.7.9a with two-step alignment strategy. Whole exome sequencing data were aligned to GENCODE hg38 reference genome using BWA-MEM algorithm. Mapped reads were further sorted by samtools.

Tumor somatic variant calling

The Genome Analysis Toolkit (GATK 4.2.0) was used for tumor somatic variant calling from whole exome sequencing data. We followed the practice guideline for somatic variant calling with GATK4 Mutect2.[15] BaseRecalibrator performed locus-based traversal operating at known sites provided by the 1000 Genome Project phase 1, dbSNP 138, Mills and 1000G gold standard indels, and hg38 known indels. Somatic variants of paired samples were called by Mutect2 accordingly. The germline mutation resource from gnomAD hg38 and a 1000G panel-of-normal were used for variant masking. Filtered variants were annotated by Ensembl-VEP 104.3.

Exitrons identification and characteristics

Exitron splicing events were identified from RNA-Seq data by ScanExitron as previously described.[9] Briefly, ScanExitron extracted uniquely mapped reads with MAPQ >50 and identifies splicing junctions. GENCODE hg38 annotated intron regions were removed. Novel junctions with canonical splicing sites located within exons were identified as exitron-splicing events. ScanExitron calculated the ratio of exitron spliced reads to overall reads across exitron regions and defined it as percent of spliced-out (PSO) value to represent the efficiency of exitron splicing. Exitrons with at least three supporting reads and PSO >0.05 were eligible for analysis. The difference of exitron burden between tumor and normal groups was measured by Wilcoxon signed rank test. Exitrons spliced in three or more normal samples, considered normal splicing events, were excluded from tumor exitrons. The rest of tumor exitrons were defined as gastric cancer-specific exitrons (GCSE). GenVisR[16] was used to visualize variant events. A threshold of 15% exitron splicing frequency was applied to display the most frequently spliced genes.

Differential splicing analysis

A generalized linear regression model-based differential test with PSO value as dependent variable and group condition as binomial independent variable was performed to detect differential exitron splicing events as previously described.[6,9] Paired tumor and normal samples were selected for analysis because splicing variant expression from unpaired samples can be dramatically varied and affect sensitivity and statistical validity. Differentially spliced cancer driver genes listed in Cancer Gene Census hallmarks (https://cancer.sanger.ac.uk/census) were highlighted. The differential spliced genes were visualized in Manhattan plot.

Identification of frequently spliced genes

A binomial distribution was used to model the exitron splicing events along exons. The background splicing probability, p, was calculated as follow:

$$p = \frac{N_{total}}{L_{total}} \quad (1)$$

where N_{total} represented the number of exitron splicing events in all genes. L_{total} represented the total length of protein-coding exons.

Furthermore, the probability of observing k_i exitron splicing events for gene i occurred in n_i length of its exon was given by the probability mass function as follow:

$$Pr(X = k) = \frac{n_i!}{K_i!(n_i - k_i)!} p^{k_i} (1 - p)^{n_i - k_i} \quad (2)$$

For each gene, the p value was further adjusted by Benjamini-Hochberg false discovery rate (FDR). Genes with FDR $< 10^{-4}$ was eventually defined as frequently exitron-spliced genes.

Functional enrichment

The hypergeometric distribution of interested genes in gene sets derived from MSigDB Hallmark (https://www.gsea-msigdb.org/gsea/msigdb/) and Gene Ontology (GO) (http://geneontology.org/) were tested using ClusterProfiler.[17] Function enrichment was visualized in dot plot and chord plot by R package ggplot2 and circlize, respectively. The width of each chord represented the number of exitron splicing events in respective frequently exitron-spliced genes.

Tissue specific test

GCSEs contained non-tissue-specific exitrons. We extracted tumor-specific exitrons from a pan-cancer cohort and excluded those detected from stomach adenocarcinoma patients. GCSEs filtered with the rest pan-cancer tumor-specific exitrons were defined as gastric cancer tissue-specific exitrons (GCTSEs).

Pathogenic exitrons detection

Pathogenic exitrons were predicted using a gradient boosting tree model.[18] A pickled pre-trained model and

VCF file containing all unique gastric cancer-specific exitrons were used as input for CAPICE predict module. The larger the CAPICE score, the more possible the exitron is pathogenic. Therefore, we used a cutoff of 0.96 for all variants to extract the most likely pathogenic exitrons. The rest of gastric cancer-specific exitrons were considered non-pathogenic, in other word, transcriptional noise.

Expression quantification and differential expression analysis

Gene expression count was measured by featureCounts.[19] The differential expression analysis was performed using DESeq2. Batch effect was controlled by setting adjusted variable in DESeq2. Normalized TPM was batch-corrected by an improved algorithm specifically designed for RNA-Seq count data, COMBAT-Seq.[20]

Expression and mutation profile of splicing factors

The expression/mutation profile and regulatory function of splicing factors (SFs) were evaluated in 128 paired samples. A gene set containing 409 SFs was collected from two independent studies.[21,22] Expression changes of SFs within individual patient were presented as the log2 fold-change of TPM normalized gene expression between tumor and paired normal sample. Some SFs with low or undetectable expression in all samples were excluded. Mutations of SFs were identified as abovementioned. Exitron-splicing events occurred in less than 5% of tumor samples were excluded. Exitrons with standard deviation of ΔPSO exceeding 0.02 were selected.

To evaluate the interaction between expression/mutation and exitron splicing, each SF was paired with a single exitron. Kendall Correlation Coefficient (KCC) was used to assess the correlation between expression/mutation status and splicing alteration. 95% confident intervals were calculated by DescTools (https://github.com/AndriSignorell/DescTools). SF-exitron pairs with $|KCC| > 0.5$ and p value < 0.05 were considered significantly correlated. For those significantly correlated pairs, positive KCC defined SF an active regulator that promoted its respective exitron splicing, otherwise the SF was defined as a negative regulator. Exitron spliced genes with identified significant splicing regulators were searched in a putative cancer driver gene database, OncoVar.[23] Genes with driver score > 20 were considered potential cancer drivers. Linear regression analysis of significant SF-driver pairs was performed. Furthermore, a regulatory network of putative cancer drivers was visualized by R packages igraph, intergraph and ggnetwork.

Mutual exclusivity analysis

Characterization of mutual exclusivity between exitron splicing events and somatic mutations was visualized by GenVisR.[16] Exitron splicing events and somatic mutations were visualized by trackViewer.[24] To annotate mutation and exitron loci with Pfam database, we extracted sequence and performed annotation using PfamScan 1.6.

Identification of compensatory and synergistic exitrons

Exitron splicing, mutation and survival data were required for complementary exitron identification. Therefore, cross-validated survival analysis was performed in 340 patients from TCGA-STAD cohort. In mutation-free patients, compensatory exitrons could compensate oncogenic effect of a driver gene mutation, resulting in non-additive consequence ($p^{neg} < 0.05$ and $p^{pos} > 0.05$). In driver gene mutated patients, synergistic exitrons could enhance oncogenic effect of driver mutations, resulting in additive consequence ($p^{pos} < 0.05$). Exitron splicing events are relatively rare comparing to mutations. Considering the retrospective design, it is necessary to decrease false positive discovery. First, splicing group should contain at least 3 exitron events. Second, a multivariate Cox regression model implemented tumor stage as a known prognosis-associated covariate and evaluated the variance inflation factor (VIF) of tumor stage and each exitron splicing event to explore the interaction between stage and exitrons. VIF larger than 10 was considered of significant interaction. Frequently mutated gene set in gastric cancer was developed from COSMIC and examined for their interaction with exitrons. Survival analysis was performed using R package survival (https://github.com/therneau/survival), rms (https://github.com/harrelfe/rms) and survminer (https://rpkgs.datanovia.com/survminer/index.html).

Genotyping HLA class I

Following the recommended online protocol, Razers3[25] was used to extract and realign HLA gene raw reads. OptiType[26] was used for HLA class I genotyping. An in-house script was used to integrate HLA-I genotyping result for neoantigen prediction.

Neoantigen prediction

VCF files were manually reformatted by in-house script. Ensembl-VEP was used to annotate GCSEs. ScanNeo workflow[27] took annotated VCFs and HLA genotypes as input to predict neoantigens with HLA binding affinity ($IC_{50} < 500nM$).

Validation of neoantigen peptides expression in CPTAC

Whole proteome mass spectrometry (MS) data of gastric cancer samples was download from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) data portal.[12]

Target database was created as follow. Inframe exitron FASTA included ten amino acids upstream and downstream of splicing junction. Frameshift exitron FASTA included ten amino acids upstream of splicing junction and all of the downstream neopeptide sequence. Target sequences were concatenated with UniProt proteome reference. To control the false discovery rate, reverse decoy sequences were concatenated with target database. Fragpipe[28] was used to perform peptide search and quantification. FDR filtering parameters were set as: –prot 0.05 –pep 0.05 –ion 0.05. Identified peptides were annotated to proteome data and visualized in TOPPView.[29] Sashimi plot was used presented evidence of neoantigen-expressing exitron splicing events at RNA level by The Integrative Genomics Viewer (IGV 2.10.0).

Correlation of neoantigen burden and immune infiltration

TPM normalized data was deconvoluted by TIMER2.0[30] (http://timer.cistrome.org/) to estimate the proportion of immune cells. Spearman correlation analysis between neoantigen burden and immune infiltration was performed. Significance level was indicated in the Spearman correlation coefficient heatmap.

Checkpoints expression and immunotherapy response

A list of immune checkpoint molecules was literature-curated. The difference of checkpoint molecules expression/immune checkpoint blockade response between high neoantigen load and low neoantigen load groups was tested by Wilcoxon rank test. A frameshift exitron splicing event triggering a premature stop codon found in (1) first exon within first 200nts of coding sequence, (2) last exon, (3) penultimate exon within 50nts of the 3' exon junction was defined a nonsense-mediated decay (NMD)-escape event as previously described.[31]

Statistical analysis

Hypergeometric test was performed for gene set enrichment test. Wilcoxon signed ranked test was used for expression comparison between groups. Fisher's exact test was used to test mutual exclusivity between exitrons and mutations. Benjamini-Hochberg FDR was calculated for multi-comparison correction. Log-rank test and univariate/multivariate Cox regression were performed for survival analysis as described above. In multivariate Cox regression model, gender, age, GCSEs load, stage and Lauren type were taken as independent variates. Imputations for missing values were performed 134 times with 250 iteration each. Imputed datasets were combined under Rubin's rule.[32] Significant level was reported at four levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Role of funders

Funders do not have a role in study design, data collection, interpretation, data analysis or writing of report in this study.

Results

Characteristics of gastric cancer exitrons

RNA-Seq, WES and MS data were collected from ENA, TCGA, CPTAC and Nanfang Hospital cohort. Overall, transcriptome of 632 patients, whole exome of 481 patients and proteome of 77 patients were analyzed in this study. The study workflow was illustrated in Figure 1. In gastric cancer samples, exitron splicing load was significantly increased (Figure 2a), reflecting disrupted exitron splicing machinery in gastric cancer. Differential spliced exitrons were observed in various cancer driver genes. 19 variants from 11 hallmark tumor driver genes were differentially spliced in tumor samples (Figure 2b). For instance, exitron AMER1 Δ50-327 locates at the N terminus of WTX domain (Figure S1a), which interacted with β-catenin, AXIN2 and APC to functionally promote ubiquitination and degradation of β-catenin, resulting in tumor suppression.[33] Inframe deletion of WTX domain in this variant may cause loss-of-function. Exitron RECQL4 Δ318-349 locates at the center of second Sld2-like DNA binding regions (Figure S1e), which have been demonstrated to be indispensable for its DNA annealing activity.[34] Exitron SPEN Δ3419-3450 locates at the C-terminal SMRT interacting domain (Figure S1f). Its transcription regulatory function requires SMRT-binding repression domain.[35] Exitrons of other driver genes e.g. BCL9L, FOXO4 and POLQ produced frameshift variants, resulting in loss-of-function effect (Figure S1b-d). Furthermore, normal exitrons were excluded and those solely expressed in gastric cancer samples were considered as gastric cancer-specific exitrons (GCSEs). 42174 GCSEs were detected in the present study. Consistent with previous findings,[7,9] majority of GCSEs are inframe (Figure 2c). Frequency of GCSE splicing for each gene was then tested to identify frequently GCSE-spliced genes. Overall, 1546 significantly spliced genes were detected (Figure 2d and Table S2). GCSEs were frequently spliced in some well-studied tumor driver genes e.g. MUC4, TAF15, FUS, KMT2D, EWSR1 etc. It is likely that these previously overlooked GCSEs play important roles in oncogenic process. To evaluate clinical relevance of GCSEs, we examined clinical features and prognostic value of GCSEs. Male, intestinal type, stage IV and ≥ 65-year-old patients have significantly higher GCSE load (Figure 2e-h). These results indicated that GCSEs could be relevant to clinical features and prognosis. Indeed, both univariate and multivariate Cox regression analyses identified GCSE load, age, stage IV as risk factors for patient prognosis (Table 1). High GCSE load

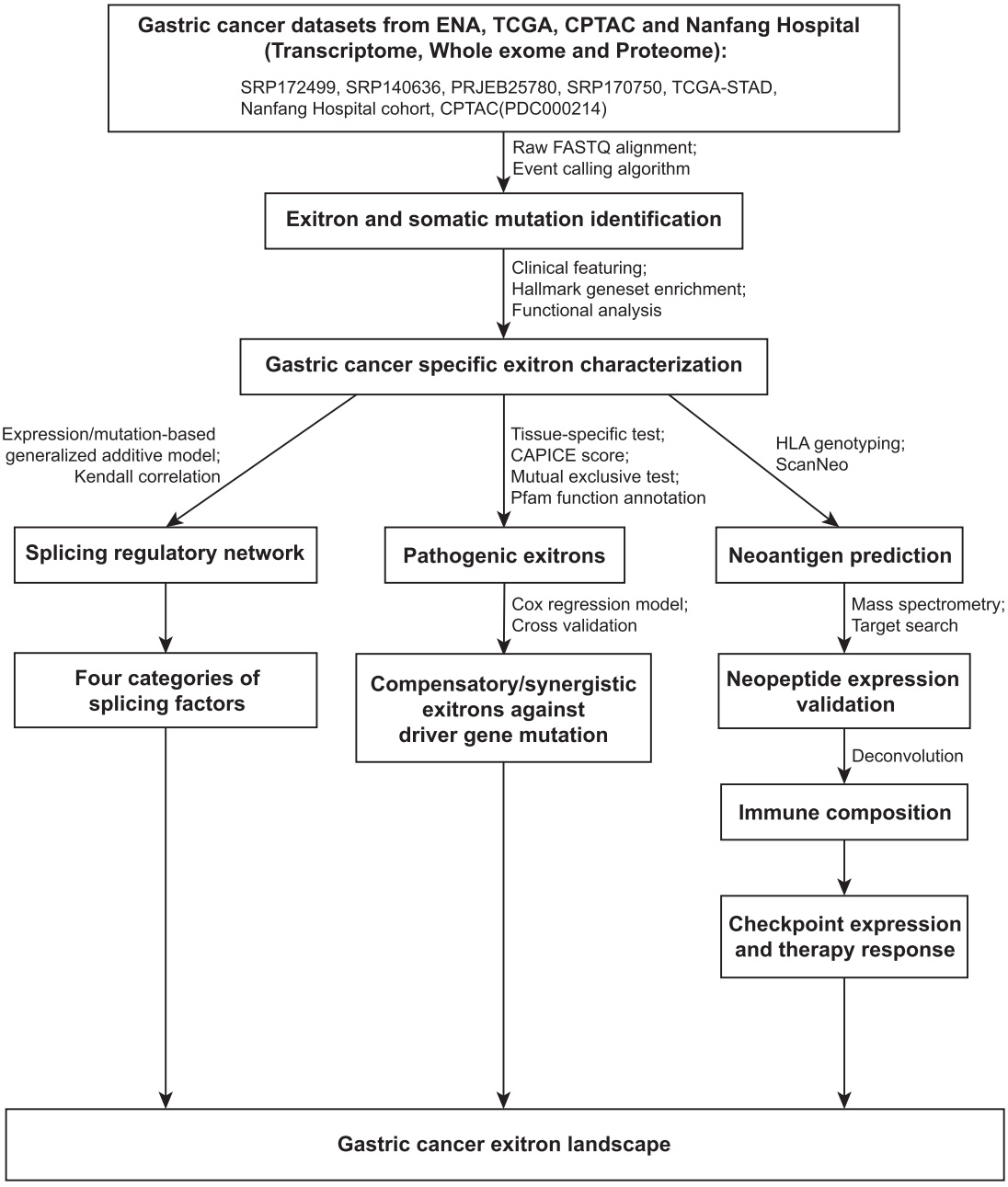

Figure 1. Workflow of the present study.

was significantly related to worse overall survival (Figure 2i). These findings highlighted the clinical relevance and unexplored importance of gastric cancer exitrons.

To understand functional impact of GCSEs, hypergeometric test was performed on Molecular Signatures Database (MSigDB) hallmark gene sets and Gene Ontology (GO). Frequently GCSE-spliced genes were significantly enriched in epithelial-mesenchymal transition (EMT, 42 genes), downregulated UV response (30 genes) and mitotic spindle (35 genes) cancer hallmark pathways (Figure 3a). A large proportion of frequently GCSE-spliced genes are not enriched in any of cancer hallmarks. Gene Ontology (GO) enrichment revealed that molecular functions are enriched in transcriptional regulation, cell junction, adhesion and cell growth (Figure S1g), which reflected the fact that disrupted splicing machinery generated a large amount of transcriptional noise and may not have any functional significance.[36] Despite GCSEs were uniquely expressed in

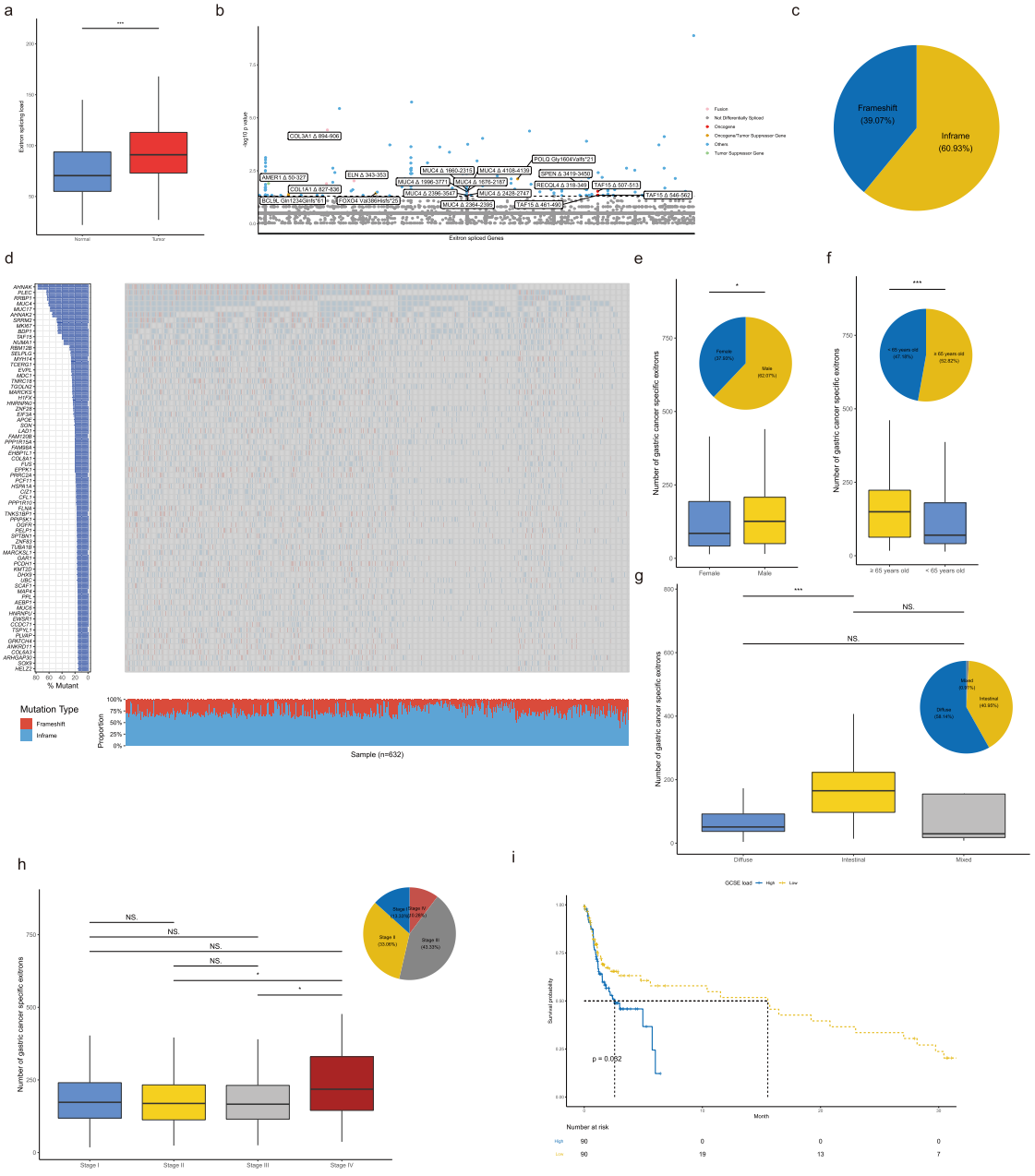

Figure 2. Characteristics and clinical features of gastric cancer exitrons.

(a) Exitron splicing load was significantly elevated in gastric cancer samples comparing to normal samples in all cohorts except for TCGA due to lack of normal samples (totally 270 tumor and 129 normal samples). p value (Wilcoxon signed rank test) (b) A linear regression model-based differential splicing analysis showed various hallmark cancer drivers contained differential spliced exitrons. 128 paired tumor/normal samples were used for test. Pink dot represented fusion gene. Red dot represented oncogene. Gold dot represented oncogene/tumor suppressor gene. Green dot represented tumor suppressor gene. Blue dot represented other genes which were not annotated as driver genes in Cancer Gene Census. (c) Proportion of frameshift and inframe GCSEs in 632 tumor samples. (d) Illustration of frequently exitron-spliced genes. Events occurred in >15% samples were showed. (e-h) GCSEs load was tested in 632 tumor samples comparing different gender (e), age (f), Lauren classification (g), and stage groups (h). Pies indicate the proportion of samples in respective groups. p value (Wilcoxon signed rank test). (i) Overall survival comparison of patients with high and low GCSE load. The first quantile was defined as high GCSE load while the last quantile was defined as low GCSE load. *p*-value (Log-rank test). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

	Univariate		Multivariate	
Variable	**Hazard ratio (95% confident interval)**	***p*-value**	**Hazard ratio (95% confident interval)**	***p*-value**
GCSE load	1.002 (1.0005-1.003)	**0.007**	1.002 (1.0003-1.003)	**0.019**
Gender				
Female	1 [Ref.]		1 [Ref.]	
Male	1.41 (0.99-2.002)	0.058	1.32 (0.92-1.90)	0.13
Age (years)	1.02 (1.008-1.04)	**0.0029**	1.03 (1.01-1.05)	**0.0004**
Stage				
I	1 [Ref.]		1 [Ref.]	
II	1.32 (0.71-2.46)	0.39	1.36 (0.73-2.57)	0.34
III	1.57 (0.87-2.84)	0.14	1.79 (0.98-3.25)	0.056
IV	3.49 (1.77-6.90)	**0.0003**	4.40 (2.20-8.82)	**0.00003**
Lauren type				
Diffuse	1 [Ref.]		1 [Ref.]	
Intestinal	0.89 0.56-1.39)	0.6	0.68 (0.42-1.10)	0.11
Mixed	NA	NA	NA	NA

***Table 1*: Cox regression analysis of overall survival in gastric cancer patients.**

gastric cancer samples, a large proportion of GCSEs were not tissue specific (Figure 3b). 19366 of 42174 GCSEs were gastric cancer tissue-specific exitrons (GCTSEs), which drew our interests to investigate the function of gastric cancer exitrons in tissue-specific manner. We excluded tumor-specific exitrons detected in all tumors except for stomach adenocarcinoma (STAD) in a pan-cancer cohort[9] to keep GCTSEs in this study. The most frequently GCTSEs-spliced genes included MUC17, AHNAK, PLEC, AHNAK2, MUC4, RRBP1, MKI67, EPPK1, BDP1, NUMA1, SRRM2, CRYBG2, TAF15 and EVPL (Figure S1h). Genes with spliced GCTSEs affected more cancer-associated functional pathways, including TNFα-NFκB, p53, estrogen response, interferon α and myc targets (Figure 3c). Furthermore, a pre-trained gradient boosting tree model was used to predict pathogenicity of GCSEs to validate the findings of GCTSEs. Higher CAPICE score indicated the higher probability of exitron pathogenicity. A threshold of 0.96 was used to determine pathogenic GCSEs. Consistently, majority of GCSEs was identified as transcriptional noise. 38.55% of GCSEs were identified as pathogenic exitrons (Figure 3d). Pathogenic GCSEs were found in many known cancer driver genes. Functional enrichment showed highly overlapped cancer pathways distribution comparing to that of GCTSEs (Figure 3e). Additionally, GO analysis demonstrated a significant enrichment of both pathogenic GCSEs and GCTSEs in true pathogenic signals rather than transcriptional noise (Figure S1i, j). We also evaluated the prognostic value of GCTSEs. 103 prognostic GCTSE-spliced genes were found (Table S3). For instance, GCTSE-spliced BCOR, as a known tumor suppressor gene, showed worse prognosis (Figure S1k). Taken together, these findings demonstrated aberrant exitron splicing has clinical importance and contributes to oncogenic process in gastric cancer.

Mutual exclusivity reveals oncogenic compensatory and synergistic exitrons

Apart from structure change induced driver effect, alternative splicing variants can be mutually exclusive with driver mutations and represent independent oncogenic process.[5,9] Despite aberrant exitron splicing is closely related to oncogenic process in gastric cancer, it remains elusive if GCSEs underlie complementary mechanism against mutations. To address this question, we evaluated mutual exclusivity of exitrons at gene and protein levels, respectively. Frequency of mutations and exitron splicing in each gene were analyzed. Driver genes like TP53, ARID1A, KMT2C, PIK3CA were the most frequently mutated genes in gastric cancer according to COSMIC database. Very low splicing frequency was detected in these genes. On the contrary, frequently spliced genes like TAF15, EWSR1, FUS had low mutation frequency (Figure 4a). Mutual exclusivity for most driver genes among individual patient was observed as well (Figure 4b). Furthermore, mutual exclusivity for event loci within genes was detected. For instance, two exitrons RECQL4 V198Pfs*43 and RECQL4 Δ 318−349 locate at Sld2-like DNA binding region of RECQL4, whereas nearly all of the mutations locate at DEAH box domain or its C-terminal region (Figure 4c). Two exitrons SPEN Δ3419−3450 and V3419Sfs*58 locates at C-terminal SMRT interaction domain of SPEN. Loss-of-function frameshift mutations of SPEN largely locate at other regions (Figure 4d). Previously, KMT2D was found to be infrequently spliced in pan-cancer cohort with significant mutual exclusivity.[9] In the present

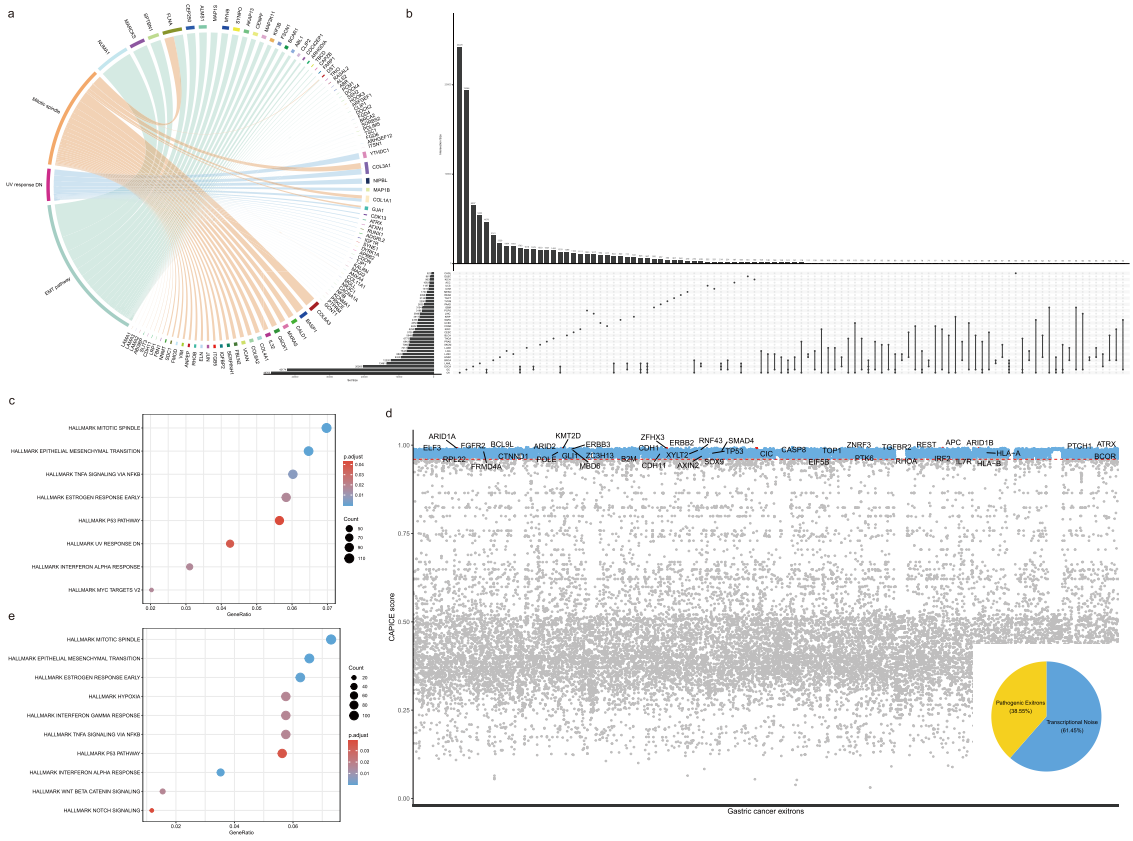

Figure 3. Functional assay of gastric cancer exitrons.

(a) Frequently GCSEs-spliced genes were enriched in MSigDB hallmark pathways. (b) intersection between GCSEs and tumor-specific exitrons in TCGA 32 tumor cohorts. GC, gastric cancer; LAML, Acute Myeloid Leukemia; ACC, Adrenocortical carcinoma; BLCA, Bladder Urothelial Carcinoma; LGG, Brain Lower Grade Glioma; BRCA, Breast invasive carcinoma; CESC, Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL, Cholangiocarcinoma; LCML, Chronic Myelogenous Leukemia; COAD, Colon adenocarcinoma; ESCA, Esophageal carcinoma; GBM, Glioblastoma multiforme; HNSC, Head and Neck squamous cell carcinoma; KICH, Kidney Chromophobe; KIRC, Kidney renal clear cell carcinoma; KIRP, Kidney renal papillary cell carcinoma; LIHC, Liver hepatocellular carcinoma; LUAD, Lung adenocarcinoma; LUSC, Lung squamous cell carcinoma; DLBC, Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; MESO, Mesothelioma; OV, Ovarian serous cystadenocarcinoma; PAAD, Pancreatic adenocarcinoma; PCPG, Pheochromocytoma and Paraganglioma; PRAD, Prostate adenocarcinoma; READ, Rectum adenocarcinoma; SARC, Sarcoma; SKCM, Skin Cutaneous Melanoma; TGCT, Testicular Germ Cell Tumors; THYM, Thymoma; THCA, Thyroid carcinoma; UCS, Uterine Carcinosarcoma; UCEC, Uterine Corpus Endometrial Carcinoma; UVM, Uveal Melanoma. (c) MSigDB hallmark enrichment analysis of gastric cancer tissue-specific exitrons. Adjusted *p*-value (Benjamini-Hochberg false discovery rate) (d) Pathogenicity prediction of GCSEs. CAPICE score ranged from zero to one. The larger the score, the more possible the pathogenicity of GCSE is. A threshold of 0.96 was used to define pathogenic exitrons. Blue dot represented pathogenic exitrons in non-cancer driver genes. Red dash line represented the 0.96 CAPICE score cutoff. Red dot represented pathogenic exitrons in cancer driver genes. Driver genes with pathogenic exitrons were labeled. A pie chart showed the proportion of pathogenic exitrons and transcriptional noise. (e) MSigDB hallmark enrichment analysis of pathogenic GCSEs. Adjusted *p*-value (Benjamini-Hochberg false discovery rate). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study, KMT2D has relatively high splicing frequency with significant mutual exclusivity. Somatic mutations were dispersed along the entire gene and more than half of exitrons overlapped with mutations (Figure 4e). Due to lack of protein structure, we are unable to map these exitrons to reveal the structural alteration in a higher dimension and directly predict their functions.

To understand if functional impact of exitrons was different from somatic mutations, we evaluate mutual exclusivity of protein function based on annotated protein database. Exitrons and mutations loci were annotated with Pfam database to highlight mutation/exitron-affected protein domains. Most mutations and exitrons were identified within Pfam annotated domains and families (Figure 4f). Mutations are less likely to occur in disordered and repeat regions. Furthermore, analyzing mutation/exitron load of each annotated protein domain revealed functional exclusivity. Mutation-

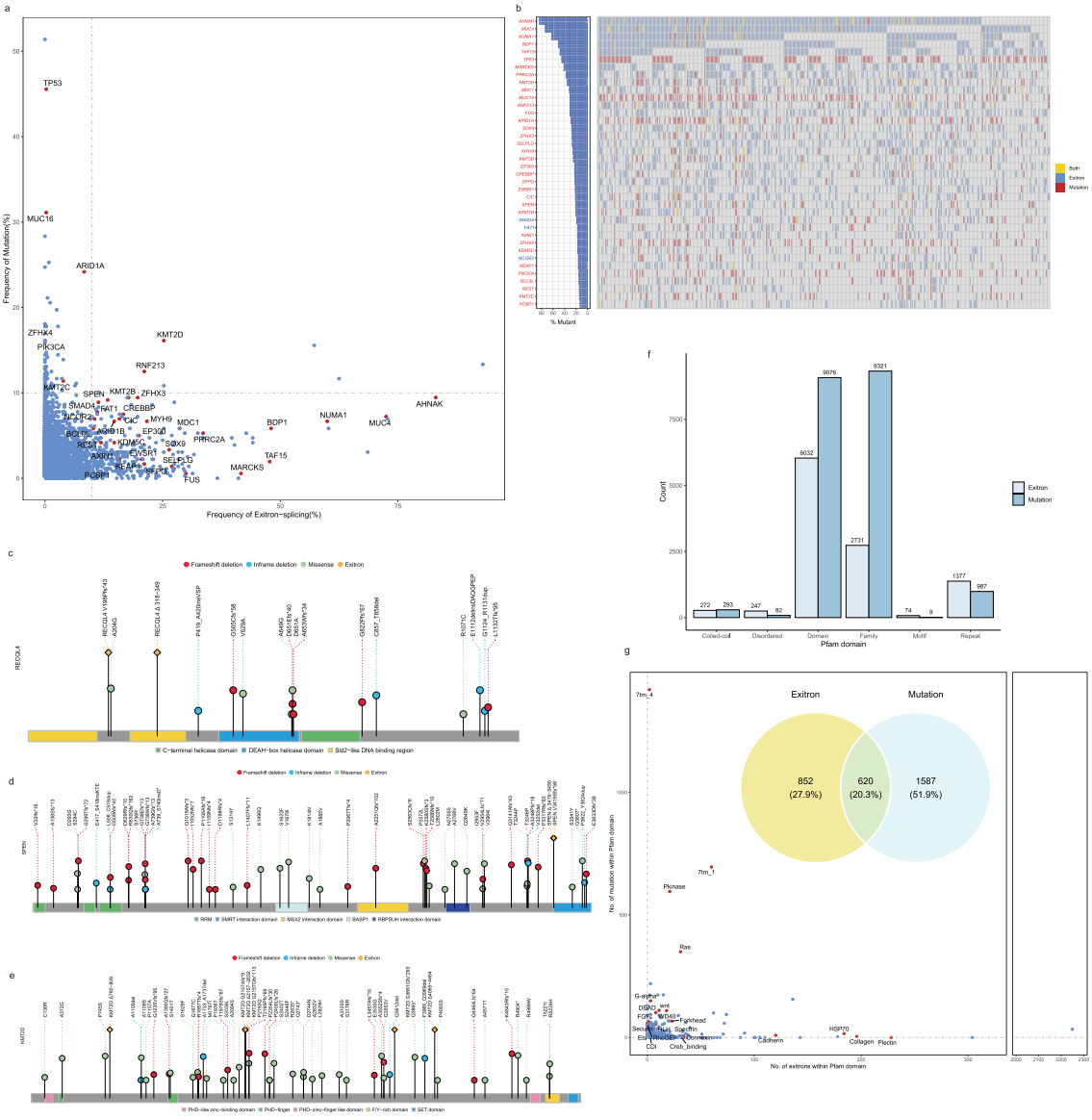

Figure 4. Mutual exclusion of gastric cancer exitrons with somatic mutations.

(a) The frequency of exitron splicing and mutations was visualized in scatter plot. Dash line represented 10% threshold. Genes of interest were highlighted in red. (b) The event occurrence of interest genes was visualized. Fisher's exact test was used to calculate the mutual exclusion of sample distribution between exitrons and somatic mutations. Significant exclusion was highlighted in red. *p*-value (Fisher's exact test). (c-e) Lolliplot indicating somatic mutations and exitrons in RECQL4, SPEN and KMT2D. (f) Barplot showed total amount of mutations/exitrons in different Pfam annotated categories. (g) Scatterplot showed the number of exitrons/mutations identified in Pfam domains. Pie chart showed the proportion of exitron/mutation-affected Pfam domains. Interested Pfam domains were highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

annotated Pfam domains and exitron-annotated Pfam domains have 620 overlaps, representing 20.3% of all annotated domains. Consistent with our previous findings, exitrons were highly enriched in functional domains related to cell junction, migration and growth (Figure 4g). Mutations on the other hand were more likely to be enriched in domains related to signal transduction and transcription regulation. In other cases, exitrons and mutations share common cancer-associated domains e.g. protein kinase, G-protein coupled receptor, Ras protein family, Wnt signaling, Forkhead. Taken together, we found that a group of exitrons specifically

affected proteins related to adhesion, migration and growth in gastric cancer. Another group of exitrons have functional impact on cancer-associated domains together with mutations, suggesting exitrons play both unique and complementary roles exclusive from mutations in gastric cancer.

Exitrons with driver effect could be under positive selection during oncogenic process, distinguishing from non-driver events.[37] To distinguish unique driver exitrons, we performed univariate Cox regression to identify prognostic exitrons. Six and four exitrons were found to be risk or protective predictors in gastric cancer, respectively (Figure 5a). For instance, the splicing of exitron SRRM2 Δ516-603 affected integrity of N-terminal transcriptional regulator domain and predicted worse prognosis in gastric cancer patients (Figure 5b). Considering that exitrons can also produce complementary effect against mutations during oncogenesis, we cross-validated prognostic exitrons in two complementary subpopulations, namely driver gene mutation positive ($Driver^{pos}$) or negative ($Driver^{neg}$) patients. Compensatory exitrons could predict $Driver^{neg}$ prognosis rather than $Driver^{pos}$ due to non-additive consequence under the same pathway. Synergistic exitrons could predict prognosis of $Driver^{pos}$ because synergistic enhancement resulted in additive consequence. Regarding to 20 highly mutated driver genes in gastric cancer, 21 compensatory exitrons and 24 synergistic exitrons were identified by multivariate Cox regression model (Table 2). most complementary exitrons were found to interact with TP53 mutation. Seven synergistic exitrons interact with $LRP1B^{pos}$ subpopulation. Two synergistic exitrons interact with $CDH1^{pos}$ subpopulation. One synergistic exitron interact with APC^{pos} subpopulation. Some well-studied cancer-associated genes including MUC4, TAF15, AXIN1, APOE and YTHDF2 were present in the list. Recently, mechanism of p53-mediated speckle association underscored the important role of nuclear speckle in cancer-associated transcription regulation.[38] It was worth noting that two exitrons in nuclear speckle formation gene, SRRM2, were categorized as both compensatory and synergistic exitrons against TP53 mutation. Exitron SRRM2 Δ1119-1147 found within C-terminal disordered region of SRRM2 (Figure S3) was identified as compensatory exitron of TP53 (Figure 5c). Exitron SRRM2 Δ516-603 on the contrary was identified as synergistic exitron of TP53 (Figure 5d). C-terminal region is indispensable for nuclear speckle formation,[39] suggesting that SRRM2 exitrons may play dual roles in p53-mediated regulation. Intriguingly, some unique prognostic exitrons were also identified as synergistic exitrons (e.g. SRRM2 Δ516-603 and CCDC85B Δ33-48). These exitrons were likely to drive oncogenic process, especially in TP53-dependent manner. Overall, we identified driver exitrons that may underlie unique or complementary mechanism in gastric cancer.

Splicing regulatory network unraveled splicing factors targeting aberrant exitron splicing

Our findings demonstrated a downstream effect of exitrons in oncogenic process of gastric cancer. It is still unclear if any upstream factor participated in it as well. Dysfunction of *trans*-acting splicing factors can greatly affect RNA splicing efficiency and produce aberrant transcripts with distinct downstream effect.[5,6] In the present study, 409 splicing factor genes were literature-curated from two independent studies[21,22] (Figure S2a). Expression of splicing factors was largely disrupted. But the expression change was not closely related to mutations. Moreover, splicing factors/individuals were not clustered by mutation load, suggesting disrupted expression and mutation in splicing factors could be different sources for aberrant exitron splicing (Figure 6a).

To address this question, we paired splicing factors with exitrons, creating two large datasets containing 112221 mutation-exitron pairs and 129745 expression-exitron pairs for Kendall correlation analysis. Kendall correlation coefficient (KCC) was used to measure the correlation between these pairs. Among expression-exitron pairs, 709 active regulators and 540 negative regulators were detected. Among mutation-exitron pairs, 339 active regulators and 293 negative regulators were detected. Gastric cancer potential driver genes (Census score > 20) were extracted from OncoVar database[23] and the top regulators were highlighted (Figure S2b, c). Furthermore, the interaction between mutation and expression of splicing factor-exitron pairs were evaluated. Majority of GCSEs were solely affected either by aberrant splicing factor expression or mutation. 214 normal exitron pairs and 17 GCSE pairs were affected by both mutation and expression change (Table S4). Splicing regulators can be classified to four categories: splicing/gain-of-function, promoting splicing/loss-of-function, suppress splicing/gain-of-function and suppress splicing/loss-of-function (Figure 6b). Positive expression correlation indicated increasing expression of regulator promoted exitron splicing, while positive mutation correlation indicated mutated regulator gained exitron splicing function.

Next, we focused on splicing regulators for tumor driver genes. Expression-based regulatory network discovered 239 splicing factors regulating 364 exitrons in 11 putative driver genes (Figure 6c and Table S5). Mutation-based regulatory network discovered 110 mutated splicing factors regulating 149 exitrons in 8 putative driver genes (Figure 6d and Table S6). No common regulators were detected in both two networks because most exitrons were solely affected either by aberrant expression or mutation as previously described (Figure 6b). It was worth noting that unique regulators for driver genes were commonly observed. Only a few splicing factors were found to regulate exitron splicing of multiple driver genes. Therefore, exitron splicing mechanism could be target-centered rather than

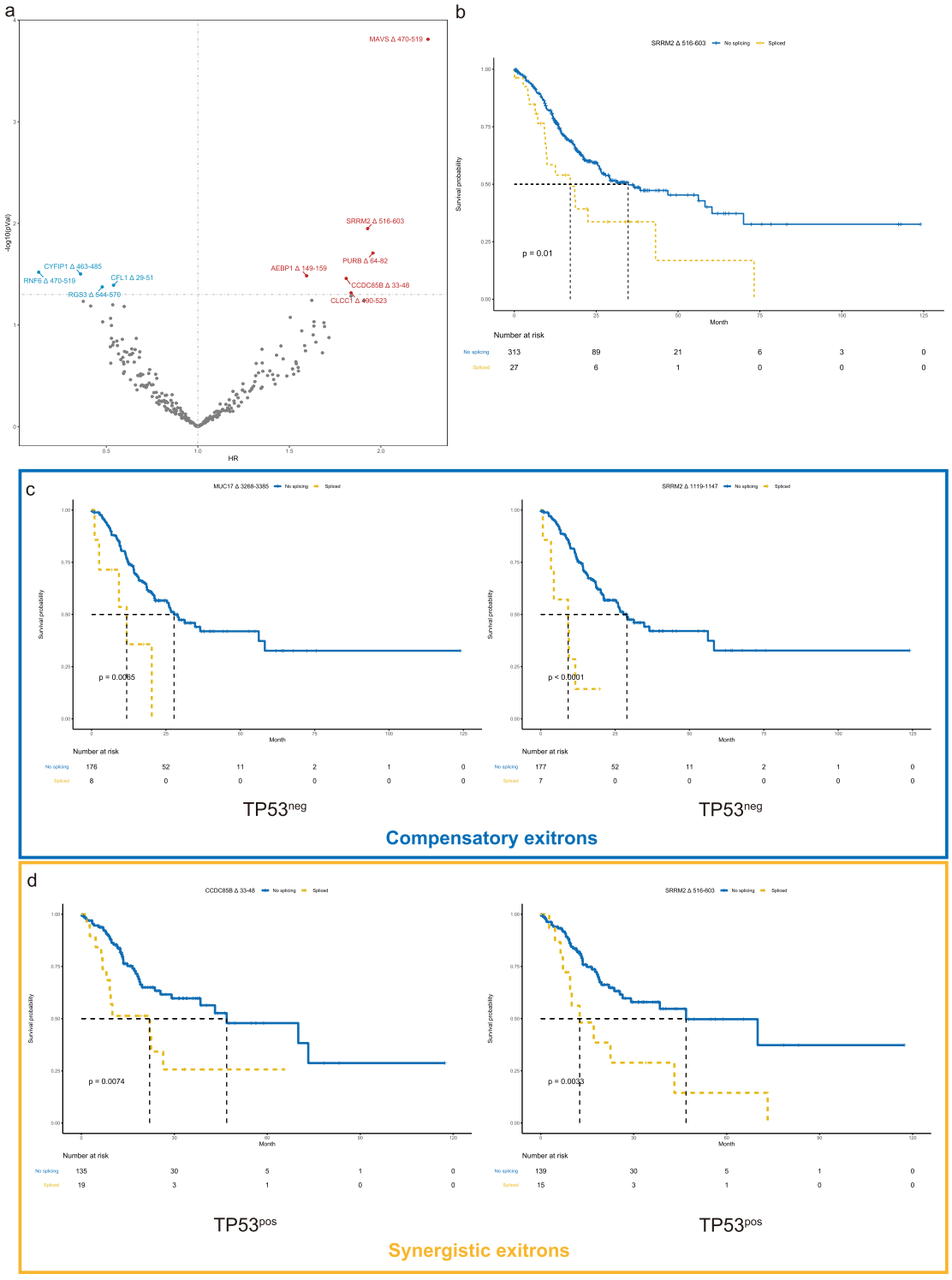

Figure 5. Identification of prognostic and complementary exitrons.

(a) Univariate Cox regression model was used to test prognostic exitrons. Red dot represented risk exitrons. Blue dot represented protective exitrons. HR, hazard ratio. *p*-value (Univariate Cox regression). (b) Splicing of exitron SRRM2 Δ516-603 resulted in worse prognosis among all patients. *p*-value (Logrank test). (c) Splicing of compensatory exitrons MUC17 Δ 3268-3385 and SRRM2 Δ 1119-1147 showed worse prognosis in TP53 mutation-free patients. *p*-value (Logrank test). (d) Splicing of synergistic exitrons CCDC85B Δ 33-48 and SRRM2 Δ 516-603 showed worse prognosis in mutated-TP53 patients. Survival analysis was performed in 340 patients. *p*-value (Logrank test). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Chromosome	Start	End	Spliced gene	Cancer driver gene
Compensatory exitrons				
chr11	62518888	62519019	AHNAK	TP53
chr14	1.05E+08	1.05E+08	AHNAK2	TP53
chr16	9092072	9092137	C16orf72	TP53
chr11	65856093	65856161	CFL1	TP53
chr1	1.54E+08	1.54E+08	CHTOP	TP53
chr6	73517849	73517890	EEF1A1	TP53
chr9	1.29E+08	1.29E+08	ENDOG	TP53
chr1	1.7E+08	1.7E+08	F5	TP53
chr3	1.29E+08	1.29E+08	H1FX	TP53
chr11	1.19E+08	1.19E+08	HYOU1	TP53
chr7	1.01E+08	1.01E+08	MUC17	TP53
chr19	50223287	50223334	MYH14	TP53
chr8	1.44E+08	1.44E+08	PLEC	TP53
chr12	11267474	11267851	PRB3	TP53
chr12	11267663	11267851	PRB3	TP53
chr7	44885103	44885159	PURB	TP53
chr16	2763884	2763970	SRRM2	TP53
chr3	1.01E+08	1.01E+08	TFG	TP53
chr12	49185097	49185384	TUBA1A	TP53
chr12	49128204	49128257	TUBA1B	TP53
chr19	23361396	23361479	ZNF91	TP53
Synergistic exitrons				
chr11	62520071	62529880	AHNAK	TP53
chr19	44908645	44908719	APOE	LRP1B
chr16	288155	288193	AXIN1	LRP1B
chr5	71510492	71510656	BDP1	TP53
chr11	65890880	65890927	CCDC85B	TP53
chr10	1.19E+08	1.19E+08	EIF3A	LRP1B
chr15	32730782	32730877	GREM1	TP53
chr19	35267258	35267352	LSR	TP53
chr7	1.01E+08	1.01E+08	MUC17	CDH1
chr7	1.01E+08	1.01E+08	MUC17	LRP1B
chr3	1.96E+08	1.96E+08	MUC4	CDH1
chr11	1017426	1018097	MUC6	TP53
chr8	1.44E+08	1.44E+08	PLEC	TP53
chr19	17365453	17365618	PLVAP	TP53
chr7	44801328	44801389	PPIA	LRP1B
chr3	51993773	51993820	RPL29	TP53
chr11	75566536	75566631	SERPINH1	TP53
chr17	81720292	81720365	SLC25A10	TP53
chr16	2762073	2762336	SRRM2	TP53
chr17	35844573	35844749	TAF15	APC
chr1	42700968	42700991	YBX1	LRP1B
chr4	68337202	68337237	YTHDC1	LRP1B
chr1	28743160	28743192	YTHDF2	TP53
chr19	57859546	57859797	ZNF587	TP53

***Table 2*: Compensatory and synergistic exitrons interacted with gastric cancer frequently mutated driver genes.**

regulator-centered. Taken together, splicing regulatory networks provided important information for unraveling new targets associated with stomach oncogenesis. Combining our findings in upstream regulatory network and downstream pathogenic effect, we provided a full picture of gastric cancer exitrons and important information for genetic screening and functional validation in the near future.

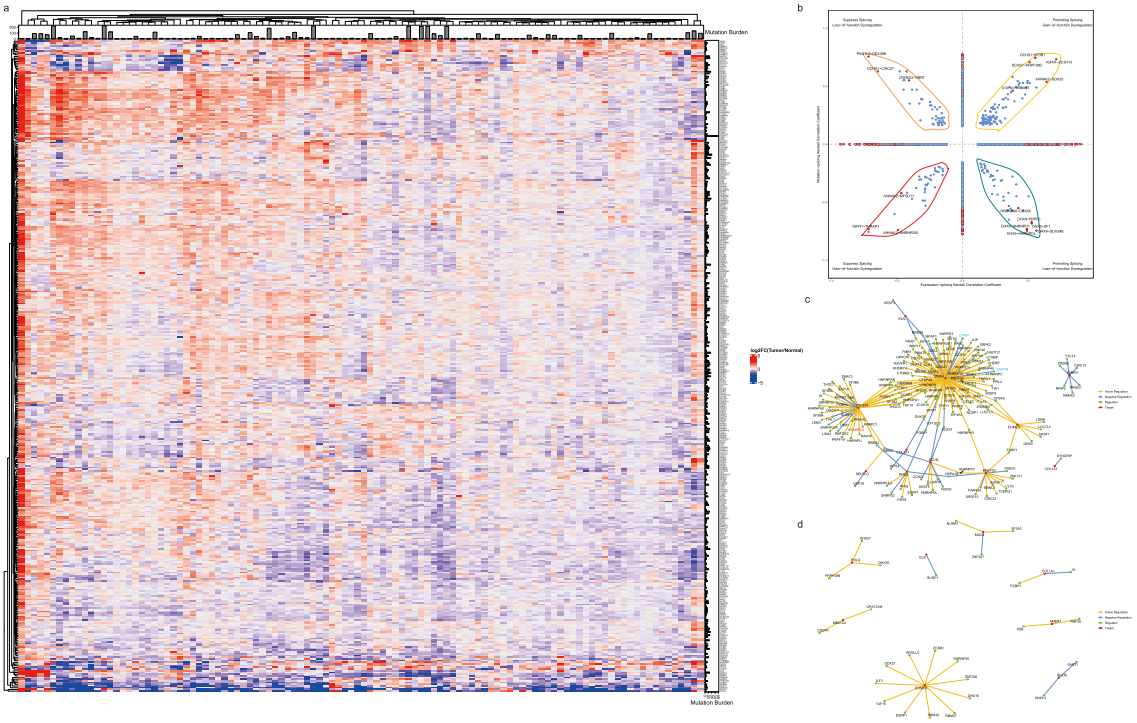

Figure 6. Upstream exitron splicing regulatory network.

(a) Heatmap revealed expression change of splicing factors in paired gastric cancer samples. Top panel represented the mutation burden in each patient. Right panel represented the mutation burden of each splicing factor among all patients. (b) Expression and mutation profile of splicing factors revealed functionally distinct groups of splicing factors regulating respective exitron splicing. Kendall correlation coefficient (KCC) was calculated for splicing factor-exitron pairs. Pairs with p value <0.05 was showed. Red dot represented splicing factor-GCSE pairs. (c) Regulatory network of putative gastric cancer drivers based on expression profile of splicing factors. Differential expression of regulators at bulk-level were also measured. Gene symbols of upregulated regulators were highlighted in red. Gene symbols of downregulated regulators were highlighted in blue. (d) Regulatory network of putative gastric cancer drivers based on mutation profile of splicing factors. In (c) and (d), red dot represented putative cancer gene nodes. Green dot represented splicing factor nodes. Gold line represented active regulation for respective pairs (KCC >0.5). Blue line represented negative regulation for respective pairs (KCC >-0.5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Exitron-derived neoantigen peptides expressed in gastric cancer

Aberrant splicing not only produce functional transcripts with driver effect but also non-functional transcripts with passenger effect. These neopeptides constituted of a large pool for tumor neoantigens.[6] On the one hand, neoantigens could be used to predict immunotherapy response in various cancer types.[40] On the other hand, neopeptides could be developed as cancer-specific vaccination targets.[41] In the present study, we found that majority of GCSEs were transcriptional noise, which could be a neopeptide pool to improve diagnosis and treatment strategies in gastric cancer. Therefore, we investigated the immunogenicity and clinical applicability of gastric cancer exitrons. Immune composition for all patients were estimated by deconvolution algorithm.[30] Majority of infiltrated immune components are M2 macrophages and resting $CD4^+$ memory T cells (Figure 7a). Frameshift transcripts have much higher propensity of immunogenicity than inframe transcripts. Following ScanNeo[27] neoantigen prediction, most predicted neopeptides were derived from frameshift GCSEs. GCSEs-derived neoantigen burden was then calculated. The burden of GCSEs-derived neoantigens had strong positive correlation with regulatory T cells, follicular T helper cells, M0 macrophages and negative correlation with plasma cells and eosinophils (Figure 7b). These results suggested a dormant microenvironment may counteract increasing immunogenicity in gastric cancer.

To validate the expression of predicted neoantigens, whole proteome mass spectrometry data of 77 samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) was analyzed. Six CPTAC-confirmed peptides expressed in 10 patients (Table 3). One peptide recurrently expressed in pt15, pt25, pt26, pt45 (Figure 7c), and one peptide recurrently expressed in pt49 and pt57. Due to lack of immunopeptidomics data,

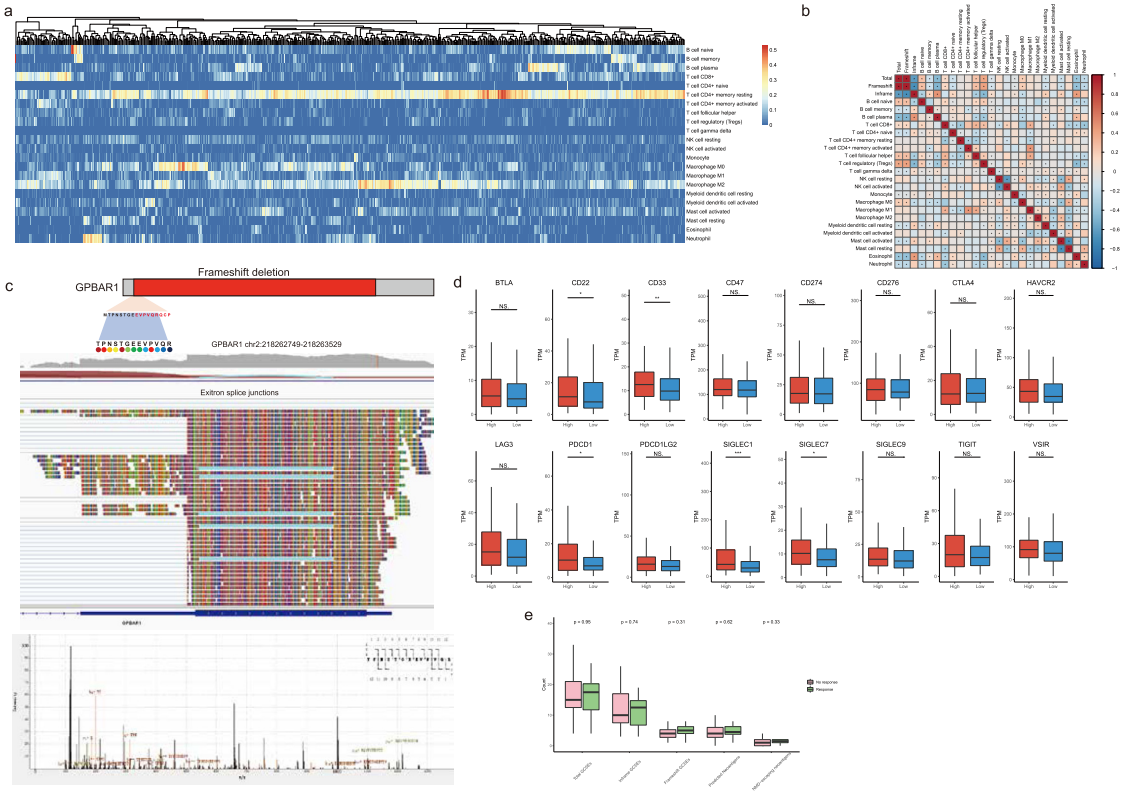

Figure 7. Immunogenicity and diagnostic potential of GCSEs.

(a) Estimated immune composition using deconvolution algorithm CIBERSORT. (b) Spearman correlation between predicted GCSEs-derived neoantigen burden and proportion of different immune component. Asterisk represented significant correlation ($p < 0.05$). *p*-value (Spearman correlation test). (c) Expression of neoantigen peptides TPNSTGEEVPVQR derived from GPBAR1 frameshift exitron splicing was confirmed by proteome mass spectrometry. (d) Expression difference of immune checkpoint molecules. *p*-value (Wilcoxon signed rank test). (e) Exitron-related biomarkers with response of immune checkpoint inhibitor. *p*-value (Wilcoxon signed rank test).

Peptide Search	Gene	Genome Position	Patient
TPNSTGEEVPVQR	GPBAR1	chr2:218262749-218263529	pt15
TPNSTGEEVPVQR	GPBAR1	chr2:218262749-218263529	pt25
TPNSTGEEVPVQR	GPBAR1	chr2:218262749-218263529	pt26
QGDGEQSAGGGPGR	FOXQ1	chr6:1312917-1312980	pt29
EEAAAGEHAGLMVTGGR	APOBR	chr16:28497455-28497799	pt33
TPNSTGEEVPVQR	GPBAR1	chr2:218262749-218263529	pt45
PGSNGNPGPPPAGNTGAPGS	COL3A1	chr2:189004001-189004036	pt49
PGSNGNPGPPPAGNTGAPGS	COL3A1	chr2:189004001-189004036	pt57
QAGECLTVLPDGAACR	ATP10A	chr15:25679498-25679598	pt59
SVEMGSVNEAYR	IGFN1	chr1:201210343-201210450	pt74

***Table 3*: Validate the expression of exitron-derived neoantigen peptides in CPTAC gastric cancer dataset.**

we were unable to validate whether exitron-derived neoantigens could be presented to major histocompatibility complex.

The expression level of immune checkpoint molecules was considered one of the indications for immune checkpoint inhibitor therapy.[42] We compared the expression of 16 literature-curated immune checkpoint molecules between high and low neoantigen load patients. The expression of PD-1, Siglec1, Siglec2, Siglec3 and Siglec7 was significantly higher in high neoantigen load group (Figure 7d). It is possible that patients with higher GCSE-derived neoantigen load

have better response for immune checkpoint blockade. Mutation-derived neoantigen burden is a valuable biomarker for immune checkpoint inhibitor treatment response.[40] Recently, growing evidence showed that nonsense-mediated mRNA decay (NMD) could repress expression of neoantigens in cancer. NMD-escape variants were associated with clinical benefit to immune checkpoint inhibitors and increased anti-tumor immunogenicity.[31,43] Therefore, we tested if different exitron biomarkers could distinguish responders from nonresponders in Kim's cohort.[11] Unfortunately, GCSE load, frameshift GCSEs, neoantigen burden and NMD-escaping neoantigen burden did not have statistical significance (Figure 7e).

Discussion

Exitron splicing load in gastric cancer cohort ranks 3rd among all the other cancers.[9] But characteristics and function of gastric cancer exitrons remains poorly understood. In the present study, we illustrated a comprehensive multidimensional landscape of gastric cancer exitrons in a large population.

Exitrons were aberrantly spliced in gastric cancer. Gastric cancer-specific exitron was significantly increased in male, elderly (≥ 65-year-old), stage IV and intestinal type patients. Increased GCSE load was related to poor prognosis. Functional assay demonstrated that GCSEs affected epithelial-mesenchymal transition (EMT), mitotic spindle and downregulated UV response. Epithelial-mesenchymal transition (EMT) was one of the most important mechanisms for epithelial-origin tumorigenesis.[44] We demonstrated that preferential enrichment of exitrons in epithelial-related genes and pathways is a unique feature in gastric cancer. From one aspect, these genes usually contain a large exon, in which exitrons are more likely to be found as previously described.[7] From another aspect, specific enrichment represents the unique function of exitrons and probably reflects an undiscovered mechanism involving in gastric cancer pathology. Tissue-specific analysis and pathogenic prediction of GCSEs were performed to validate a confident subset of pathogenic gastric cancer exitrons. Genes with pathogenic exitron splicing uniquely enriched in additional cancer-associated pathways including p53, TNFα-NFκB, estrogen response and interferon α response pathways, together suggesting that pathogenic exitrons were likely to contribute to oncogenic process of gastric cancer by affecting respective cancer pathways.

Exitron shows mutual exclusivity with mutations at individual, gene and protein level. We demonstrated that exitrons can have unique function or complementary function with tumor driver genes. A prognostic exitron cross validation was performed to identify compensatory or synergistic exitrons against tumor driver gene mutations. One candidate gene and its exitrons has drawn our attention. SRRM2 and SON are indispensable paired components for nuclear speckle formation.[39] Recent study demonstrated that p53-mediated nuclear speckle association regulated the RNA amount of p53 targets for downstream transcription tuning.[38] These studies imply that SRRM2 could be of particular importance in oncogenesis. In the present study, we identified compensatory exitron SRRM2 Δ1119-1147 and synergistic exitron SRRM2 Δ516-603 against TP53 mutation. SRRM2 Δ1119-1147 is located at C-terminal disordered region. SRRM2 Δ516-603 overlaps with transcription regulator domain. A previous report showed that nuclear speckle formation was disrupted with expression of C-terminal truncated SRRM2 where N terminus remained intact.[39] Our findings suggested SRRM2 Δ516-603 and SRRM2 Δ1119-1147 could have distinct functional role in p53-dependent manner. More interestingly, splicing of SRRM2 Δ516-603 could predict prognosis among all patients. Hence, splicing of SRRM2 exitrons is likely to be a unexplored mechanism in oncogenic process of gastric cancer.

To understand if upstream regulatory factors elicit the downstream effect caused by aberrant exitron splicing, we established a splicing regulatory network to identify splicing factors targeting respective exitrons. Because the expression change of only a few splicing factors could be detected at bulk level (Figure S2D), individual paired test was necessary to increase the sensitivity for detection. We found that exitron splicing regulation was target-centered rather than regulator-centered. Moreover, we listed splicing regulator candidates for tumor driver gene exitrons. These findings should provide useful information for experimental screening and validation in the future.

Immune checkpoint inhibitor is a promising treatment for gastric cancer. However, clinical benefit could only be observed in specific patients.[45] Discovery of biomarkers to predict checkpoint inhibitor response is in urgent need. Currently, microsatellite status, EB virus infection, tumor mutation burden, neoantigen burden, immune gene signatures, plasma cells are demonstrated to be predictive biomarkers in solid tumors.[11,46-48] In the present study, significant increase of immune checkpoints (PD-1, Siglec1, Siglec2, Siglec3 and Siglec7) was detected in high GCSE-derived neoantigen load patients. Intriguingly, high GCSE-derived neoantigen load was found to be correlated with Siglecs expression. We do not know if exitrons have any interactions with Siglecs or Siglecs-targeted therapy. Further investigation may provide more evidence. Despite expression difference of PD-1 was detected, we failed to distinguish responders from nonresponders using any of GCSEs-related biomarkers. A possible explanation was that data collected from Kim's cohort was small sample size and the transcriptome sequencing is at a relatively low resolution. These could lead to insufficient calling of aberrant splicing events. Previously, a significant association

between exitron-derived neoantigen burden and clinical benefit could only be observed in clear cell renal cell carcinoma but not in melanoma.[9] Therefore, we were unable to draw any conclusion so far. The predictive value of exitrons in gastric cancer remains elusive and a larger, better-designed, high-resolution study is needed.

The present study has several limitations. First, functions of most exitron-spliced genes in cancer are overlooked. Considering a large proportion of exitron splicing might be transcriptional noise, it will be a challenge to screen and validate cancer driver genes from this gene set. Moreover, the present study is an *in-silico* analysis. Therefore, experimental screening and validation is necessary to uncover the relevant mechanism. Second, it is better to identify exitrons using RNA-based, full-length sequencing strategy due to potential artifacts found recently.[49] But this technique is not easily performed. Third, most samples in CPTAC gastric cancer cohort are derived from diffuse-type early-onset gastric cancer patients. Considering that exitron splicing load is much higher in intestinal patients, the selection bias in this cohort will underrepresent the diversity and amount of exitron-derived neoantigens.

Taken together, we illustrated a multidimensional comprehensive landscape of gastric cancer exitrons. GCSEs splicing had functional impact on oncogenic process and clinical importance. Pathogenic gastric cancer exitrons demonstrated unique function and epithelial function preference. More importantly, we identified complementary GCSEs against TP53 mutations and revealed potential targets for previously undiscovered oncogenic mechanism in gastric cancer. Splicing network was also established to illustrate upstream regulatory candidates eliciting downstream effect via exitron splicing. Additionally, the immunogenicity of GCSEs-derived neoantigens was validated by MS. GCSEs-derived neoantigen load could distinguish patients with different expression of PD-1 and Siglecs. But GCSEs-related biomarkers failed to predict checkpoint blockade response, possibly due to insufficient sequencing depth. A large-scale genetic screening and functional validation will be necessary to provide more solid evidence and support our findings.

Contributors

Y.H. Zhang conceptualized the study, collected data from ENA and TCGA, performed exitron detection and somatic variant call, analyzed characteristics of exitrons and mutations, performed prediction of neoantigens, searched target database and wrote the manuscript. G. T. Ye, Q.B. Yang, G.F. Zhang and B.Y. Zheng collected patient samples and performed transcriptome RNA sequencing at Nanfang Hospital. J. Yu and Y.F. Hu supervised sample collection and data processing. Y.H. Zhang, Q.B. Yang and G.X. Li verified the underlying data. G.X. Li supervised all experiments, data interpretation and revised the manuscript. G.X. Li was responsible for the decision of submission. All authors read and approved the final version of the manuscript, and ensure it is the case.

Data sharing statement

Public RNA-Seq and WES data were available at European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) under the following accession number: SRP172499, SRP140636, PRJEB25780, SRP170750. RNA-Seq raw data from Nanfang Hospital cohort is available upon reasonable request. Proteomic data (PDC000214) was available at CPTAC data portal. Processed exitrons and exitron-derived neoantigens detected in TCGA-STAD cohort was available at Mendeley via vdkpfzjjvg.1.

Declaration of interests

The authors declare no potential conflicts of interest.

Acknowledgements

We appreciated the support of following grants: the Guangdong Provincial Key Laboratory of Precision Medicine for Gastroinstestinal Cancer (2020B121201004), the Guangdong Provincial Major Talents Project (No. 2019JC05Y361) and National Natural Science Foundation of China (grant number:82172960 and 81872013).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104272.

References

1 Thrift AP, Nguyen TH. Gastric Cancer Epidemiology. *Gastrointest Endosc Clin N Am*. 2021;31(3):425–439.
2 Lott PC, Carvajal-Carmona LG. Resolving gastric cancer aetiology: an update in genetic predisposition. *Lancet Gastroenterol Hepatol*. 2018;3(12):874–883.
3 Wang K, Yuen ST, Xu J, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat Genet*. 2014;46(6):573–582.
4 Gravalos C, Jimeno A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. *Ann Oncol*. 2008;19(9):1523–1529.
5 Climente-Gonzalez H, Porta-Pardo E, Godzik A, Eyras E. The functional impact of alternative splicing in cancer. *Cell Rep*. 2017;20(9):2215–2226.
6 Kahles A, Lehmann KV, Toussaint NC, et al. Comprehensive analysis of alternative splicing across tumors from 8705 patients. *Cancer Cell*. 2018;34(2). 211-24 e6.
7 Marquez Y, Hopfler M, Ayatollahi Z, Barta A, Kalyna M. Unmasking alternative splicing inside protein-coding exons defines exitrons and their role in proteome plasticity. *Genome Res*. 2015;25(7):995–1007.
8 Yang R, Van Etten JL, Dehm SM. Indel detection from DNA and RNA sequencing data with transIndel. *BMC Genomics*. 2018;19(1):270.
9 Wang TY, Liu Q, Ren Y, et al. A pan-cancer transcriptome analysis of exitron splicing identifies novel cancer driver genes and neoepitopes. *Mol Cell*. 2021;81(10). 2246-60 e12.

10 Mendiratta G, Ke E, Aziz M, Liarakos D, Tong M, Stites EC. Cancer gene mutation frequencies for the U.S. population. *Nat Commun*. 2021;12(1):5961.
11 Kim ST, Cristescu R, Bass AJ, et al. Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gastric cancer. *Nat Med*. 2018;24(9):1449–1458.
12 Mun DG, Bhin J, Kim S, et al. Proteogenomic characterization of human early-onset gastric cancer. *Cancer Cell*. 2019;35(1). 111-24 e10.
13 Kim SK, Kim HJ, Park JL, et al. Identification of a molecular signature of prognostic subtypes in diffuse-type gastric cancer. *Gastric Cancer*. 2020;23(3):473–482.
14 Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–ii90.
15 Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43. 11 0 1- 0 33.
16 Skidmore ZL, Wagner AH, Lesurf R, et al. GenVisR: genomic visualizations in R. *Bioinformatics*. 2016;32(19):3012–3014.
17 Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*. 2021;2(3):100141.
18 Li S, van der Velde KJ, de Ridder D, et al. CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation of Clinical Exome variations. *Genome Med*. 2020;12(1):75.
19 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930.
20 Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. 2020;2(3):lqaa078.
21 Giulietti M, Piva F, D'Antonio M, et al. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res*. 2013;41(Database issue):D125–D131.
22 Seiler M, Peng S, Agrawal AA, et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Rep*. 2018;23(1). 282-96 e4.
23 Wang T, Ruan S, Zhao X, et al. OncoVar: an integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Res*. 2021;49(D1):D1289–DD301.
24 Ou J, Zhu LJ. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. *Nat Methods*. 2019;16(6):453–454.
25 Weese D, Holtgrewe M, Reinert K. RazerS 3: faster, fully sensitive read mapping. *Bioinformatics*. 2012;28(20):2592–2599.
26 Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014;30(23):3310–3316.
27 Wang TY, Wang L, Alam SK, Hoeppner LH, Yang R. ScanNeo: identifying indel-derived neoantigens using RNA-Seq data. *Bioinformatics*. 2019;35(20):4159–4161.
28 Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods*. 2017;14(5):513–520.
29 Sturm M, Kohlbacher O. TOPPView: an open-source viewer for mass spectrometry data. *J Proteome Res*. 2009;8(7):3760–3763.
30 Li T, Fu J, Zeng Z, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res*. 2020;48(W1):W509–WW14.
31 Litchfield K, Reading JL, Lim EL, et al. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. *Nat Commun*. 2020;11(1):3800.
32 Rubin DB. Multiple Imputation for Nonresponse in Surveys; 1987.
33 Major MB, Camp ND, Berndt JD, et al. Wilms tumor suppressor WTX negatively regulates WNT/beta-catenin signaling. *Science*. 2007;316(5827):1043–1046.
34 Keller H, Kiosze K, Sachsenweger J, et al. The intrinsically disordered amino-terminal region of human RecQL4: multiple DNA-binding domains confer annealing, strand exchange and G4 DNA binding. *Nucleic Acids Res*. 2014;42(20):12614–12627.
35 Shi Y, Downes M, Xie W, et al. Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. *Genes Dev*. 2001;15(9):1140–1151.
36 Melamud E, Moult J. Stochastic noise in splicing machinery. *Nucleic Acids Res*. 2009;37(14):4873–4886.
37 Martinez-Jimenez F, Muinos F, Sentis I, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;20(10):555–572.
38 Alexander KA, Cote A, Nguyen SC, et al. p53 mediates target gene association with nuclear speckles for amplified RNA expression. *Mol Cell*. 2021;81(8). 1666-81 e6.
39 Ilik IA, Malszycki M, Lubke AK, Schade C, Meierhofer D, Aktas T. SON and SRRM2 are essential for nuclear speckle formation. *Elife*. 2020;9.
40 Desrichard A, Snyder A, Chan TA. Cancer neoantigens and applications for immunotherapy. *Clin Cancer Res*. 2016;22(4):807–812.
41 Frankiw L, Baltimore D, Li G. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol*. 2019;19(11):675–687.
42 Darvin P, Toor SM, Sasidharan Nair V, Elkord E. Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp Mol Med*. 2018;50(12):1–11.
43 Supek F, Lehner B, Lindeboom RGH. To NMD or not to NMD: nonsense-mediated mRNA decay in cancer and other genetic diseases. *Trends Genet*. 2021;37(7):657–668.
44 Ribatti D, Tamma R, Annese T. Epithelial-Mesenchymal transition in cancer: a historical overview. *Transl Oncol*. 2020;13(6):100773.
45 Kono K, Nakajima S, Mimura K. Current status of immune checkpoint inhibitors for gastric cancer. *Gastric Cancer*. 2020;23(4):565–578.
46 Gibney GT, Weiner LM, Atkins MB. Predictive biomarkers for checkpoint inhibitor-based immunotherapy. *Lancet Oncol*. 2016;17(12):e542–ee51.
47 Patil NS, Nabet BY, Muller S, et al. Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. *Cancer Cell*. 2022;40(3). 289-300 e4.
48 Yarchoan M, Hopkins A, Jaffee EM. Tumor mutational burden and response rate to PD-1 inhibition. *N Engl J Med*. 2017;377(25):2500–2501.
49 Schulz L, Torres-Diz M, Cortes-Lopez M, et al. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol*. 2021;22(1):190.